TP3 | **Network reconstruction and analysis**
Janv.-Fév. 2019 | **Durée: (voir date sur le site de cours)**

Objectif: Analysis of genomic and ploidy alterations.

- Le TP peut être réalisé seul ou en binôme. Merci de bien préciser les noms et prénoms de chaque membre du binôme et de mettre en copie votre binôme lors de l'envoi des scripts et de votre rapport.
- Les résultats de vos analyses doivent être commentés et reprendre les différents notions vues en cours.
- Vous devez me faire parvenir votre projet (scripts et rapport) pour la date indiquée sur le site de cours à (*severine.affeldt@parisdescartes.fr*).
  Titre du message: [**MLDS-app. Graphes**] ou [**MLDS-fi. Graphes**]

Analysis of genomic and ploidy alterations in breast tumors

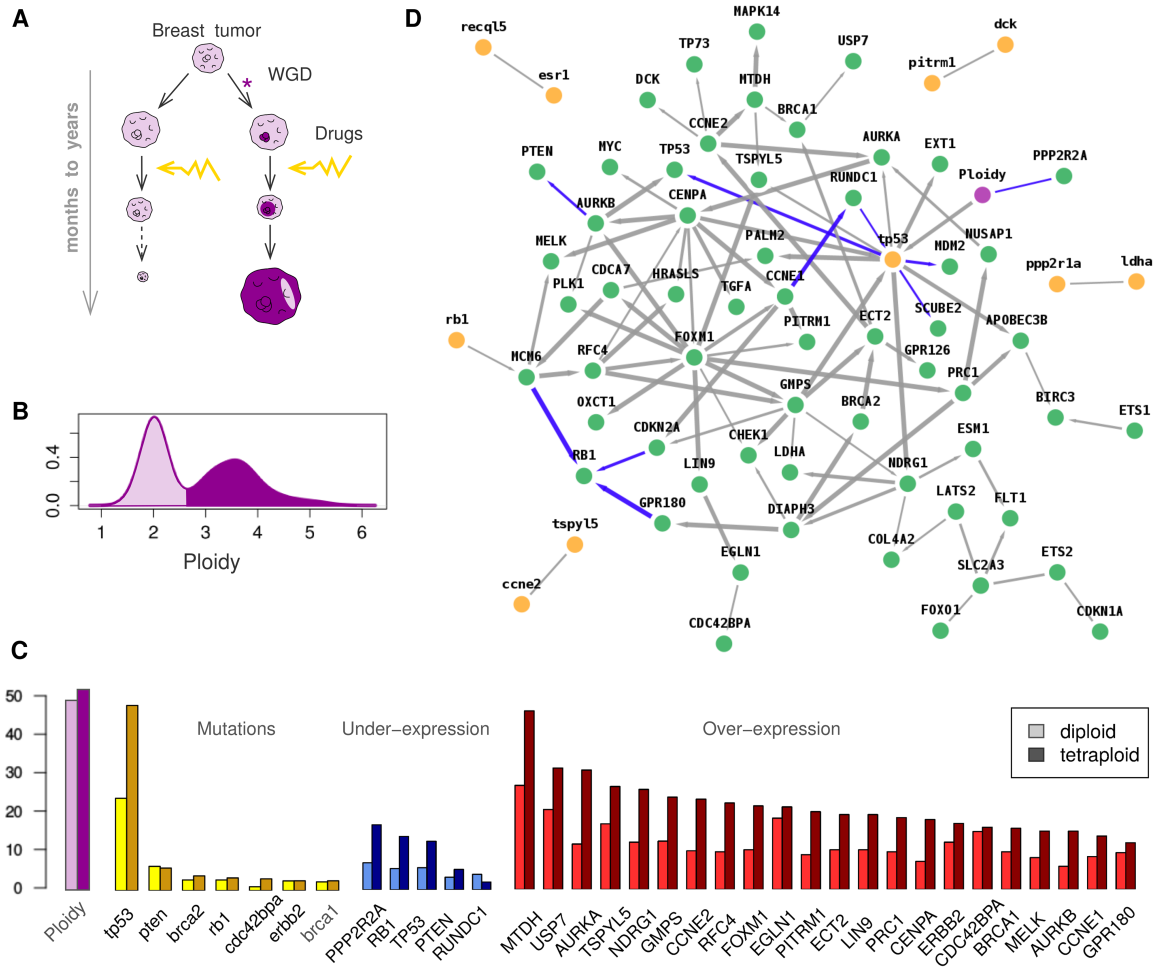
Motivations

We want to analyze genomic alterations on breast tumors from the online Catalog of Somatic Mutations in Cancer (COSMIC). The dataset, which contains 807 samples without predisposing BRCA1/2 germline muta- tions, includes somatic mutations (from whole exome sequencing) and expression level information for 91 genes. These 91 genes have been selected based on earlier studies on mutation and/or expression alterations in breast cancer, Materials and Methods. Gene non-synonymous mutation status is binarized (yes / no) and gene expression status is categorized as under-, normal- or over-expressed by the COSMIC database.

In addition to gene mutations and altered expression levels, we also integrated information on sample average ploidy, provided by the COSMIC database (release v76) and discretized the clearly bimodal ploidy distribution with ploidy ¡ 2.7 considered as diploid cells and 2.7 taken as tetraploid cells, in agreement with COSMIC convention. Among the 807 samples, 401 correspond to diploid tumoral cells and 398 to tetraploid tumoral cells (8 samples have no ploidy information).

The following figure provides the `MIIC` network reconstruction at tissue level[1]. The Figure (A) schematize the tumor development and drug resistance in the presence of tetraploid tumor cells following whole genome duplication (WGD). In (B), you can find the ploidy distribution in the 807 tumor samples. The histogram in (C) represents the genomic alterations: ploidy, mutations, normalized under-expression and over-expression changes from COSMIC database. Finally, the figure (D) is the `MIIC` Genomic alteration network obtained between average ploidy (violet), gene mutations (yellow, lower case) and under- or over-expressions (green, upper case).

[1] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005662



1. **Network reconstruction with the hill-climbing approach**

   1. From the package MIIC, load the *cosmicCancer* data using the function `data()`. Explore the dataset content, variables and dimensions.

   2. Call the hill-climbing approach from the `bnlearn` package and identify the issues related to the observational data if any. Propose a way to resolve them.

   3. Remove the samples for which at least one variable has a `NA` value. Use the function `complete.cases()` to perform this action. What proportion of the dataset did you remove? What are the concerned variables? Call again the hill-climbing approach. *NB: The 'Ploidy' variable may be interpreted as 'integer'. If necessary, convert this column into factor type. Furthermore, the variables should have at least 2 levels. Remove the column with constant variable.*

   4. Convert the hill-climbing network to an igraph object and plot the result. Some variables can have 0 degree. Do not display these variables. Put different colors for mutated genes (lower case), over/under expressed genes (upper case) and Ploidy. *NB: The graph can be dense. Consider using the function* `qgraph.layout.fruchtermanreingold` *from the* `qgraph` *package*[2]

[2] Example: https://stackoverflow.com/questions/39290909/igraph-resolving-tight-overlapping-nodes

5. Idendify the mutated genes (lower case nodes) that are significantly related to gene expression (upper case nodes). Identify also the variables related to the 'Ploidy' property. On which nodes are centered the hubs? Get top 10 nodes and edges in terms of betweenness centrality measure.

2. **Network reconstruction with the PC approach**

   1. From the package `MIIC`, load the *cosmicCancer* data using the function `data()`. Explore the dataset content, variables and dimensions.
   2. Call the `PC` approach from the `pcalg` package and identify the issues related to the observational data if any. Propose a way to resolve them.
      NB: Follow the example given in the documentation to call PC. Use the 'disc' independence test.
   3. Remove the samples for which at least one variable has a `NA` value. Use the function `complete.cases()` to perform this action. What proportion of the dataset did you remove? What are the concerned variables? Call again the PC approach.
      NB: The `Ploidy` *variable may be interpreted as 'integer'. If necessary, convert this column into factor type. Furthermore, the variables should have at least 2 levels. Remove the column with constant variable. You should also convert all the factor into their levels, with the first level being 0.*
   4. Convert the `PC` network to a `bn` object then to an `igraph` object and plot the result. Some variables can have 0 degree. Do not display these variables. Put different colors for mutated genes (lower case), over/under expressed genes (upper case) and Ploidy.
      NB: The graph can be dense. Consider using the function
      `qgraph.layout.fruchtermanreingold` *from the qgraph package*[3]
   5. Produce several graphs at different significance levels.
   6. Idendify the mutated genes (lower case nodes) that are significantly related to gene expression (upper case nodes). Identify also the variables related to the `Ploidy` property. On which nodes are centered the hubs? Get top 10 nodes and edges on terms of betweenness centrality measure.

3. **Network reconstruction with the MIIC approach**

   1. Follow the example from the documentation to build the *cosmicCancer* network with `MIIC`.
   2. Explain the arguments `confidenceShuffle` and `confidenceThreshold`. Get several networks with different values for these arguments
   3. Convert the `MIIC` network to an `igraph` object and plot the result. Some variables can have 0 degree. Do not display these variables. Put different colors for mutated genes (lower case), over/under expressed genes (upper case) and Ploidy.
      NB: The graph can be dense. Consider using the function
      `qgraph.layout.fruchtermanreingold` *from the* `qgraph` *package*[4].

[3] Example: https://stackoverflow.com/questions/39290909/igraph-resolving-tight-overlapping-nodes

[4] Example: https://stackoverflow.com/questions/39290909/igraph-resolving-tight-overlapping-nodes

   4. Use also the cytoscape fonctionality to produce a `MIIC` plot with your preferred network.

4. Idendify the mutated genes (lower case nodes) that are significantly related to gene expression (upper case nodes). Identify also the variables related to the 'Ploidy' property. On which nodes are centered the hubs? Get top 10 nodes and edges on terms of betweenness centrality measure.